How Much Do You Believe?

Nic Wilson*
Department of Computer Science
Queen Mary and Westfield College
London, England

ABSTRACT

This paper responds to a number of criticisms of Dempster-Shafer theory made by Judea Pearl. He criticises Dempster-Shafer belief for not obeying the laws of Bayesian belief; however, these laws lead to well-known problems in the face of ignorance, and seem unreasonably restrictive. It is argued that it is not reasonable to expect a measure of belief to obey Pearl's sandwich principle. The standard representation of "if-then" rules in Dempster-Shafer theory, criticised by Pearl, is justified and favorably contrasted with a conditional probability representation.

KEYWORDS: ***Dempster-Shafer theory, Bayesian probability, lower probability, conditioning, the sandwich principle, if-then rules***

1. INTRODUCTION

Judea Pearl's most challenging and stimulating paper raises a number of very important issues about belief functions and uncertain inference in general. Pearl's main criticisms can be summarized as follows:

1. Dempster–Shafer belief in proposition a can be interpreted as the probability that the evidence proves a. Pearl asks why the probability of provability should be used rather than just the probability.
2. He points out that constraints on probability functions can usually not be represented by a belief function.
3. He criticizes the conditioning in Dempster–Shafer theory.
4. Dempster–Shafer belief is shown not to obey the "sandwich principle," which Pearl argues is a fundamental principle of plausible reasoning.

Address correspondence to Nic Wilson, Department of Computer Science, Queen Mary and Westfield College, Mile End Road, London E1 4NS, England.

*This research was partly supported by the ESPRIT basic research action DRUMS (3085).

International Journal of Approximate Reasoning 1992; 6:345–365

655 Avenue of the Americas, New York, NY 10010 0888-613X/92/$05.00

5. He claims that the Dempster–Shafer representation of rules has counter-intuitive properties.

This reply first argues that the use of probability can be problematic; in particular, a single probability function is not always adequate for representing someone's degree of belief. Dempster's rule of conditioning is then contrasted, by use of an example, with Bayesian updating. The sandwich principle is examined, and it is found that measures of belief and support generally do not obey this principle, and I conclude that it certainly cannot be regarded as a general principle of plausible reasoning. It is then shown that the standard representation of rules in the Dempster–Shafer theory has a simple and natural semantics, and some major problems with other representations of rules that Pearl prefers are indicated.

2. PROBLEMS WITH PROBABILITY

Perhaps Pearl's most fundamental question is why a probability-based measure such as Dempster–Shafer belief should be used to measure belief rather than just a probability: "Shouldn't we really be concerned with the likelihood that a component is faulty as opposed to the likelihood that it can be proven faulty?" (P.6).[1] But the requirement that a measure of belief should be a probability function is very restrictive; in particular, in the case where we have very little knowledge about some proposition a, it prevents us from allocating small degrees of belief to both a and $\neg a$.

In this discussion of some problems of probability, I will focus on the Bayesian approach; maximum entropy and lower probability will be mentioned in the next section.

In the field of uncertainty in AI there has been very substantial progress in both theoretical and applications-based work on Bayesian methods (due in no small way to Pearl himself), for example, Pearl [1], Lauritzen and Spiegelhalter [2], and many papers given at recent uncertainty workshops [3]. Yet despite this important work, there are still the same familiar fundamental problems with the Bayesian approach.

2.1. Representation of Ignorance—The Frame Problem

Let a be the proposition that I live on Cowley Road, Oxford. I imagine that the reader is pretty ignorant about the truth of a. How would you construct $\Pr(a)$, your Bayesian belief in a?

First you must choose a frame Θ (of mutually exclusive and exhaustive propositions) and a subset A of Θ representing the proposition a; and then

[1] This refers to Section 6 of Pearl's paper, *Int. J. Approx. Reasoning* 4(5/6), 363–390, 1990.

presumably you would use the principle of indifference to arrive at your Bayesian belief in a. The problem is that there are any number of frames that you could choose; for example,

1. $\Theta = \{x_1, x_2\}$, $A = \{x_1\}$. The principle of indifference gives $\Pr(A)$ to be 0.5.
2. Or you could estimate that there are about 100,000 roads or streets in central England, and so pick $\Theta = \{x_1, x_2, \ldots, x_{100,000}\}$ with again $A = \{x_1\}$ and the other x_i's representing the other streets. Then the principle of indifference gives $\Pr(A) = 0.00001$.

Neither of these frames seems unreasonable, but the probability assigned to a is crucially dependent upon which frame is chosen. Sometimes there will be a clearly natural choice of frame: a unique frame such that all of its elements seem equipossible, for example, when considering the result of throwing a die, but often there will be no such "canonical" frame. A very basic requirement of an uncertainty calculus is that it should give logically equivalent propositions the same measure; the example above shows that probability fails in this requirement. One's Bayesian belief is thus a function not only of the information given and one's background knowledge, but also of a sometimes arbitrary choice of frame. An adequate measure of belief for this situation would not depend on the choice of frame; for example, Dempster–Shafer or a lower probability approach would lead to beliefs of 0 for both a and $\neg a$.

The above example suggests that Bayesian belief does not correspond to all one's intuitions about belief. Having a Bayesian belief of 0.99999 in $\neg a$ does not mean that one can be almost certain that a is false; in the above example, one would arrive at that degree of belief by choosing frame 2, but the choice could have been arbitrary, and perhaps if the same question had been asked a few seconds later the other frame would have been chosen, giving a belief of 0.5.

There is an analogous problem in the infinite frame case. For example, if one is attempting to predict tertiary protein structure (Clark et al. [4]), using a Bayesian approach requires one to produce a prior distribution on the very large dimensional space of all tertiary structures for proteins, about which we are largely ignorant (and then update this using the evidences). But if a uniform prior is used, then the probability of events in the space will depend radically on the choice of one of the infinite number of parameterizations of the space; again there is no canonical or "correct" parameterization, and with some parameterizations a given event will have prior probability almost 1, but with others almost 0.

Complete ignorance is a rather extreme case, but in a large complex frame there is almost certainly some ignorance, and the same problem occurs. A related problem occurs with likelihoods. In the Dempster–Shafer theory one can very naturally represent a piece of evidence supporting a proposition a but saying nothing extra about any subpropositions of a, or about $\neg a$, a point

made by Gordon and Shortliffe [5]. In a Bayesian approach, using a frame Θ and subset A to represent a, one would model such evidence by using a likelihood function constant over A (Pearl [6]). But in this Bayesian representation the evidence supports each $x \in A$ equally, which is not the same at all as evidence supporting A and saying nothing else about $x \in A$. A sort of "principle of indifference" is being used to split support for A equally among the singletons; to be precise, if e represents this evidence, then for each $x \in A$, $\Pr(e \mid x)$ is set to $\Pr(e \mid A)$, which is exactly what we wanted to avoid doing!

The likelihood function (as well as the prior function) is effectively a function only of the elementary propositions (the x's) and not of the whole Boolean algebra of propositions, and so it cannot represent such evidence entirely adequately.

2.2. The Propagation of Wayward Priors

The Bayesian problem with ignorant priors would perhaps not matter much if the problem disappeared as soon as evidence arrived. This, however, is not the case; unless the evidence is extremely strong, so that all but one possibility is virtually eliminated, the prior is still very important.

Bayes' rule, in odds-likelihood form is $O_e(h) = L_e(h)O(h)$, where $O(h)$ is the prior odds for hypothesis h, $O_e(h)$ the posterior odds after receiving evidence e, and $L_e(h) \equiv \Pr(e \mid h)/\Pr(e \mid \neg h)$ the likelihood ratio for evidence e.

An immediate consequence of this equation is that if an expert modifies his prior odds for some h by a factor k, that is, he replaces his prior odds $O(h)$ by $O'(h) = kO(h)$, then his posterior odds for h will be changed by the same factor, that is, $O'_e(h) = kO_e(h)$. Typically the prior probability for h is an estimate of the proportion of the time the expert expects h to occur in those circumstances (e.g., what proportion of patients who come into the surgery have disease h). With a large frame it can very easily happen that an expert "misjudges" his priors by an order of magnitude, and sometimes considerably more, especially if a principle of indifference had to be used. This means that the posterior odds can be trusted up to, at best, an order of magnitude. If the evidence very strongly supports hypothesis h leading to posterior probability 0.999, then an order of magnitude difference either way in the prior odds will only lead to a range of posterior probabilities [0.99, 0.9999] so the prior makes little difference, but if we get a posterior probability of 0.5 it should be remembered that one order of magnitude variation in prior odds would lead to posterior probabilities between (roughly) 0.1 and 0.9.

The logical conclusion seems, then, to be that a Bayesian should allow for a range of priors leading to a range of posteriors, perhaps using the infimum of these values as the measure of belief. This is then a *lower probability*

approach, and this measure of belief, as shown in Section 5.4, spoils the sandwich principle.

3. INCOMPLETENESS

Pearl argues (P.2) that one usually cannot represent a set of probabilistic constraints by a belief function. This is not surprising because belief functions are very special lower probability functions, ones with a nonnegative Möbius inverse (mass function). It is clear from the papers of Dempster and Shafer that the Dempster–Shafer theory was never intended for this purpose. However, belief functions do represent a very natural form of "incomplete knowledge" —when we have a probability function on a space Ω and wish to extend this via a compatibility relation to another space Θ (Shafer [7]).

The problems of representation of ignorance by a probability distribution discussed in the preceding section were caused by not having information to determine a single probability function, which is "incompleteness" in Pearl's sense. Representation of evidence supporting a nonelementary proposition, also discussed in Section 2, is another very important form of "incomplete knowledge," which Dempster–Shafer theory deals with in a very natural way.

A major weakness of this section of Pearl's paper is the omission of any discussion of alternative approaches to representation of incomplete knowledge. I will briefly discuss two of these other approaches.

Bayesians either seem to refuse to acknowledge the existence of incomplete knowledge, expecting their expert to be able to produce a complete set of probabilities no matter how ignorant they are, or use the principle of indifference or a generalization, maximum entropy. The very serious problems caused by dependence on the choice of frame have been mentioned in Section 2.

3.1. Maximum Entropy

Pearl makes a very misleading statement (P.2) when he states that the maximum entropy approach (Jaynes [8]) allows "the missing probabilities [to] be recovered." If the constraints on a probability function P allow one to deduce only that $P \in \mathscr{P}$, a family of probability functions, then any rule that picks a single probability function $P_1 \in \mathscr{P}$ is almost certain to produce the wrong one and may well not even be close. Maximum entropy is such a rule, which has nice mathematical properties (and also some odd ones) (Paris and Vencovská [9, 10])—but it would be a great mistake to consider that the single probability function it produces could summarize $\mathscr{P}$ at all adequately.

Take $\Theta = \{x, y\}$ and the constraint on P (on Θ), $0.05 \leq P(x) \leq 0.45$, and let $\mathscr{P}$ be the set of probability functions on Θ compatible with this constraint. We might decide to pick $P_1 \in \mathscr{P}$ with $P_1(x) = 0.25$, but it must

be remembered that this P_1 is no more *correct* than any other member of $\mathscr{P}$—we certainly have not recovered the missing probability.

In fact, maximum entropy picks probability function P_2 with $P_2(x) = 0.45$, which is a long way from some members of $\mathscr{P}$. What maximum entropy does is to pick the probability function $P \in \mathscr{P}$ that is closest, in a particular sense, to P_0, the uniform distribution. However, if there is more than one reasonable choice of frame (a common situation), then P_0 loses its special status, and so finding the P closest to P_0 is of very dubious value.

3.2. Upper and Lower Probability

Upper and lower probability approaches (e.g., Smith [11], Nilsson [12], van der Gaag [13]) seem a natural way to deal with incomplete probabilistic information. There do appear to be complexity problems, though, and the resulting intervals can be disappointingly weak. A possible solution to this latter problem might be to make extra independence assumptions.

4. DEMPSTER OR BAYESIAN UPDATING?

Pearl criticizes the method used in the Dempster–Shafer theory for updating with certain evidence, Dempster's rule of conditioning, and others have also expressed misgivings (e.g., Kyburg [14], Fagin and Halpern [15]). Considering a belief function Bel as a lower probability, the infimum of a particular family $\mathscr{P}$ of probability functions [$\text{Bel}(A) = \inf_{P \in \mathscr{P}} P(A)$], leads to an alternative way of conditioning on certain evidence B: The conditioned function is defined to be the infimum of the conditioned family, $\text{Bel}(A \,\|\, B) = \inf_{P \in \mathscr{P}} P(A \mid B)$. Kyburg calls such updating *Bayesian*. Pearl refers to it as "straight conditioning" (P.3.1) and "FH conditioning" (P.3.3). It has also been considered by Dempster [16], De Campos et al. [17], Fagin and Halpern [15], Jaffray [18]. Dempster's rule of conditioning differs from Bayesian conditioning by making some natural assumptions. This is illustrated in the following example.

A doctor states that a certain person is anemic (proposition a). Model the doctor as a randomly reliable source or sometimes reliable truth machine (Shafer and Tversky [19], Shafer [7]) so that he has two modes, reliable and unreliable, and if he is reliable then what he says is true, but if he is unreliable then we don't know whether his statement is true or not. Suppose our subjective probability that he is reliable is 0.95.

We then observe that the patient is a Bayesian (proposition b), which we treat as certain evidence. Let E_1 represent the event that the doctor is reliable, and let $\Omega_1 = \{E_1, \neg E_1\}$; let E_2 be the event that the observation is reliable; let $\Omega_2 = \{E_2, \neg E_2\}$ and $\Omega = \Omega_1 \times \Omega_2$. By the definition of a randomly

reliable source, E_1 implies a and E_2 implies b. Let the frame $\Theta = \{a \wedge b, a \wedge \neg b, \neg a \wedge b, \neg a \wedge \neg b\}$.

Consider the situation before we observe his Bayesianism. Using a Dempster–Shafer approach, we construct a probability function P_1 on Ω_1 and extend it to a belief function over Θ. Our information about the reliability of the doctor means that we should take $P_1(E_1) = 0.95$, which leads to a belief function Bel_1 over Θ with $\text{Bel}_1(a) = 0.95$.

A lower probability approach would treat the doctor's evidence as a constraint on the probability functions on Θ: $\Pr(a) \geq 0.95$, leading to the same belief function Bel_1 on Θ. The two approaches so far are essentially equivalent; the difference comes when we condition on the observation.

In a Dempster–Shafer approach, we construct a probability function P on Ω summarizing the two evidences and then extend this to a belief function Bel over Θ. Dempster's rule of conditioning makes the "irrelevance" assumption that since the two evidences are not conflicting, the probability of E_1 should not be affected by learning the second evidence, that is, $P(E_1) = P_1(E_1) = 0.95$. This leads to $\text{Bel}_1(a \mid b)$, which is by definition $\text{Bel}(a)$, equalling 0.95. Bayesian conditioning does not make this assumption; instead, $\text{Bel}_1(a \| b) \equiv P_*(a \mid b) = 0$.[2]

Dempster's conditioning leads to a more intuitive result in this case by making a sensible default assumption. There will, however, be times when other assumptions may be more appropriate, leading to a different probability function P on Ω and in turn to a different belief function on Θ, and sometimes we should perhaps allow for a family of probability functions on Ω and hence a family of belief functions on Θ, as Pearl suggests (P.5).

5. THE SANDWICH PRINCIPLE

The sandwich principle, suggested by Pearl to be a fundamental principle of plausible reasoning, may be paraphrased as follows:

> If Belf is a measure of "degree of belief," then, for any given situation and for any propositions a and b, it should be the case that $\text{Belf}(a)$ is between, or equal to one of, $\text{Belf}(a \mid b)$ and $\text{Belf}(a \mid \neg b)$, where $\text{Belf}(a \mid b)$ is the conditional belief in a given b.

At first sight, aside from the fact that terms *degree of belief* and *conditional belief* have not be defined, this seems plausible; no one will argue that it is a mathematical truth for probabilities measuring ratios of frequencies. But why should it apply to a measure of belief?

[2] Also, $P_*(a \mid \neg b) = 0$, spoiling the sandwich principle.

With a Dempster–Shafer approach, since receiving either a piece of certain evidence b or certain evidence $\neg b$ reduces uncertainty in our knowledge, it can happen that either will increase our belief in a. Conversely, both b and $\neg b$ may contradict (different) evidences that support a, and so sometimes it is the case that $\text{Belf}(a) > \text{Belf}(a \mid b)$, $\text{Belf}(a \mid \neg b)$, also violating the sandwich principle.

Although Pearl criticizes Dempster–Shafer belief for not obeying the sandwich principle, he also faces difficulties. Pearl's version of the Bayesian theory involves the consideration of a family $\mathscr{P}$ of probability functions (P.5). To make sense of the sandwich principle for this situation we have to define a function Belf measuring belief, based on the information given by $\mathscr{P}$. Perhaps the most obvious measure is lower probability, but this violates the sandwich principle (see Section 5.4). However, using maximum entropy to pick Belf means that it can depend on arbitrary choice of frame; also, considering maximum entropy on a conditioned family of probability functions leads to a violation of the sandwich principle.

For a set of probability functions $\mathscr{P}$, let $\mathscr{P}_{\text{ME}}$ represent the maximum entropy probability function for $\mathscr{P}$. Let $\Theta = \{a, \neg a\} \times \{b, \neg b\}$, and let $\mathscr{P} = \{P \text{ on } \Theta\colon\ P(a) = 0.2\}$. Conditioning $\mathscr{P}$ on new evidence b gives $\mathscr{P}_b = \{P_b\colon\ P \in \mathscr{P}\}$, which is the set of all probability functions on $\{a, \neg a\}$. Then $\mathscr{P}_{\text{ME}}(a) = 0.2$, but $(\mathscr{P}_b)_{\text{ME}}(a) = (\mathscr{P}_{\neg b})_{\text{ME}}(a) = 0.5$, thus violating the sandwich principle.

One could also define the conditioned belief as just ordinary conditioning of the probability function $\mathscr{P}_{\text{ME}}$, which, of course, does obey the sandwich principle. But with this definition, belief is not a function just of the family of probability functions; when we update with certain evidence b, our new measure of belief in a no longer has the same relationship with $\mathscr{P}_b$ as our initial belief in a had with $\mathscr{P}$.

Even if someone were to accept the sandwich principle as a principle of plausible reasoning, it doesn't necessarily mean that they shouldn't use Dempster–Shafer theory. With each belief function is associated the set of compatible probability functions (see, e.g., Dempster [16] and Section 4 of my reply to Shafer's paper). So making a decision (using Dempster's approach) requires considering a family of probability functions, just as Pearl must; it does not matter if we call the infimum of this family a "measure of belief" or not (though I happen to find it intuitive to do so).

In Section 5.1 it is argued that a function measuring the *measure of support given by the evidence* should spoil the sandwich principle, and also that the sandwich principle is incompatible with an adequate representation of ignorance. The likelihood ratio is rather like such a measure, so it is not surprising that it too spoils the sandwich principle; but since, given a state of prior ignorance, Bayesian probability can generally be considered as just normalized likelihoods, why, then, does Bayesian probability not spoil the sandwich

principle? This is explored in Section 5.2. Section 5.3 shows how the sandwich principle fails to hold in another probabilistic framework, and Section 5.4 that lower probability also spoils the sandwich principle. Finally, in 5.5 an example is given in which the use of a measure of belief that does not spoil the sandwich principle will give a very counterintuitive result and so questions whether any measure of belief should obey the sandwich principle.

5.1. Measures of Support Violate the Sandwich Principle

Let Supp be a function from an algebra of propositions to [0, 1] that is intended to represent the support that the evidence we have received gives to propositions. Supp is also meant to be such that if proposition a is established with certainty, then $\text{Supp}(a) = 1$, and if a is not supported at all by any evidence, then $\text{Supp}(a) = 0$.

Let a and c be logically independent propositions and b the proposition $a \Leftrightarrow c$. (Some may be more comfortable with the problem expressed in terms of a frame Θ; for example, we could set $\Theta = \{u, v, w, x, y, z\}$, represent a by the set $\{u, z\}$, b by $\{u, v, w\}$, and c by $\{u, x, y\}$.) Suppose we were initially ignorant about whether a is true or not but now have two sources of evidence, the first telling us that c is true, the second telling us that $\neg c$ is true. Since neither evidence lends any support to a, $\text{Supp}(a) = 0$.

Let $\text{Supp}(a \mid b)$ represent the support for a given that I learn new evidence b. Given b, the first evidence supports a, so $\text{Supp}(a \mid b)$ should certainly be greater than 0, and similarly, given $\neg b$, the second evidence supports a, so $\text{Supp}(a \mid \neg b) > 0$. Therefore a measure of the *support that evidence gives* should not obey the sandwich principle.

However, since we're in a state of prior ignorance, this result should transfer to a measure of belief as well—in the absence of prior information, a measure of belief should depend only on the evidence, that is, the information summarized by the support function, and belief should thus be a strictly monotonic function of support. This means that since $\text{Supp}(a)$ is less than both $\text{Supp}(a \mid b)$ and $\text{Supp}(a \mid \neg b)$, $\text{Belf}(a)$ should be less than $\text{Belf}(a \mid b)$ and $\text{Belf}(a \mid \neg b)$, thus spoiling the sandwich principle for belief also. The sandwich principle is thus incompatible with an adequate representation of ignorance.

Pearl's argument seems to be that, since given b we have some belief in a, and given $\neg b$ we have some belief in a, we therefore must have some belief in a without knowing either b or $\neg b$. But neither of the evidences in any way supports a, and thus we are completely ignorant about the truth of a, so why should we have a nonzero belief in a?

A Bayesian cannot have $\text{Belf}(a) = 0$ because this means for him essentially that a cannot happen; he might look at the frame Θ and then decide that $\text{Belf}(a)$ should be $1/3$; but the choice of frame was arbitrary; if instead we had

set

$$\Theta = \{u, v_1, \ldots, v_{12}, w_1, \ldots, w_{12}, x_1, \ldots, x_{12}, y_1, \ldots, y_{12}, z\}$$

representing b by $\{u, v_1, \ldots, v_{12}, w_1, \ldots, w_{12}\}$, etc., that Bayesian would then have calculated Belf(a) to be 0.04.

5.2. Likelihoods and the Sandwich Principle

The likelihood ratio $L_e(a)$ is defined as $\Pr(e \mid a)/\Pr(e \mid \neg a)$ for hypothesis a and new evidence e. We wish to condition on proposition b. Pearl, rightly I believe, allows b to be treated as either a piece of new evidence or an added assumption (P.3.2).

Treating b as a new piece of evidence means that the conditional likelihood ratio would be defined as $L_{e \wedge b}(a)$. It turns out that this type of conditioning does obey the sandwich principle. If, on the other hand, we consider b as an assumption or context so that the conditional likelihood ratio $L_e(a \mid b)$ is defined by $\Pr(e \mid a, b)/\Pr(e \mid \neg a, b)$, then the sandwich principle is spoiled. To show this, one need only use an example demonstrating Simpson's paradox (Simpson [20], Pearl [1, p. 495]) or take $\Pr(a) = \Pr(b) = 1/2$, $\Pr(b \mid a) = 7/8$, $\Pr(e \mid a, b) = \Pr(e \mid \neg a, b) = 9/10$, $\Pr(e \mid a, \neg b) = \Pr(e \mid \neg a, \neg b) = 1/10$. This leads to $L_e(a \mid b) = L_e(a \mid \neg b) = 1$, but $L_e(a) = 4$.

Now suppose we know only the values of the likelihoods $L_e(a \mid b) = L_e(a \mid \neg b) = 1$, $L_e(a) = 4$ and we are ignorant about the priors. By Bayes' rule, $O_e(a) = L_e(a)O(a)$, $O_e(a \mid b) = L_e(a \mid b)O(a \mid b)$, and $O_e(a \mid \neg b) = L_e(a \mid \neg b)O(a \mid \neg b)$, where $O(a)$ is the prior odds for a, $O_e(a)$ is the posterior odds for a after learning evidence e, etc.

If a uniform prior is used, then the posterior odds are proportional to the likelihood ratios. The likelihoods do not obey the sandwich principle, so how can the posteriors?

Consider $O(a)$. The most sensible value one can give for this is 1, that is, $P(a) = 1/2$, using the principle of indifference. This gives the posterior $O_e(a) = 4$, so $\Pr_e(a) = 0.8$. Similarly, we can consider just $O(a \mid b)$; again the most sensible value that this can be given is 1, leading to $\Pr_e(a \mid b) = 0.5$. Using the same argument, the best value we can give to $\Pr_e(a \mid \neg b)$ is 0.5. Thus the best values we can individually give to $\Pr_e(a)$, $\Pr_e(a \mid b)$, and $\Pr_e(a \mid \neg b)$ do not obey the sandwich principle.

Obviously, these values taken together are not probabilistically coherent. Thus, any posterior probability function must differ in at least one of these three values, and so must be, in some sense, suboptimal; in particular, a uniform prior is inconsistent with the likelihoods. Any prior with $\Pr(a \mid b) = \Pr(a \mid \neg b)$ turns out to be probabilistically incoherent, so we must set $\Pr(a \mid b) > \Pr(a \mid \neg b)$ or $\Pr(a \mid b) < \Pr(a \mid \neg b)$ despite the fact that the information

given in the problem is symmetric with respect to b and $\neg b$. Therefore, no single prior probability function is adequate for this problem.

Now consider the following situation. We are given just the likelihoods $L_e(a \mid b) = L_e(a \mid \neg b) = 1$ and are ignorant about priors. Using symmetry, or the principle of insufficient reasons, gives $\Pr(a \mid b) = \Pr(a \mid \neg b) = 0.5$, leading to $\Pr_e(a \mid b) = \Pr_e(a \mid \neg b) = \Pr_e(a) = 0.5$, which then implies that $L_e(a) = 1$, an unreasonable conclusion. The supposedly noninformative priors allow one to *deduce* that the sandwich principle holds for these likelihoods, which is essentially the same as assuming that Simpson's paradox does not occur.

5.3. Random Probabilities and the Sandwich Principle

At my favorite restaurant there are two possible starters, avocado or prawn cocktail, and each may be served with a roll and butter or not. I, however, have no choice in what I receive, it being decided by the waiter, who uses the following compound experiment.

First, numbers $p, q, r \in [0, 1]$ are picked randomly with some unknown distribution. Let $s = pq + (1 - p)r$. All I know about the random generation process is that the expected value of s, $E[s]$, is 0.2, and $E[q] = E[r] = 0.5$. [There is nothing contradictory in this; for example, the random generation process may be picking $(p, q, r) = (1/8, 9/10, 1/10)$ with probability 0.5, and picking $(p, q, r) = (7/8, 1/10, 9/10)$ *with probability* 0.5.]

Then I am given a roll and butter (b) with probability p, and

1. If b, then I'm given avocado (a) with probability q.

2. If $\neg b$, then I'm given avocado with probability r.

I'm interested in the values q, r and s, i.e., the probabilities $\Pr(a \mid b)$, $\Pr(a \mid \neg b)$ and $\Pr(a)$ in the random starter generation. But the best value I can give for q, and r, is 0.5, and the best value I can give for s is 0.2, again spoiling the sandwich principle.

5.4. Lower Probability and the Sandwich Principle

One week, to supplement my income I decided to venture into the catering industry. I made, with the help of many friends, a very large number of sandwiches, half with brown bread, half with white bread, which I intended to sell to various shops.

The shops require that at least 70% of the sandwiches they receive be of top quality (A grade), and I spent considerable time checking that 75% of the sandwiches I made were A grade. Unfortunately, I then found out that all the shops I had been intending to sell to wanted either only brown bread sandwiches or only white bread sandwiches.

I pick one of my sandwiches at random. Let a represent that it is of A grade quality, and b represent that it is made of brown bread. $\Pr(a) = 0.75$ and

$\Pr(b) = 0.5$. The bounds for $\Pr(a \mid b)$ are $0.5 \leq \Pr(a \mid b) \leq 1$, and those for $\Pr(a \mid \neg b)$ are $0.5 \leq \Pr(a \mid \neg b) \leq 1$. Thus if $\Pr_*$ is the lower probability function, then $\Pr_*(a) = 0.75$, $\Pr_*(a \mid b) = \Pr_*(a \mid \neg b) = 0.5$, spoiling the sandwich principle and showing that we cannot be sure that the collection of brown bread sandwiches and the collection of white bread sandwiches are up to the required quality.

Consider replacing the use of lower probability by a measure Belf that obeys the sandwich principle, such as probability. It is natural to set $\text{Belf}(a) = 0.75$, and, since the problem is symmetric with respect to b and $\neg b$, we should have that $\text{Belf}(a \mid b) = \text{Belf}(a \mid \neg b)$. But then by the sandwich principle $\text{Belf}(a \mid b) = 0.75$, suggesting that the brown bread sandwiches are up to the required standard—which is surely the wrong answer.

It is very easy to find examples where lower probability spoils the sandwich principle much more dramatically; for example, see Section 4. The main point of this example was to show that spoiling the sandwich principle is no flaw of lower probability; in this example it is the lower probability in which we are interested, and to give a satisfactory answer it must spoil the sandwich principle.

5.5. The Philippe, Pearl, and Mary Problem

A patient has exactly one of a set of diseases $\Theta = \{a_1, \ldots, a_{1000}\}$, and Mary is attempting to deduce information about which. A certain test on the patient is taken; let c mean that the result of the test is positive and $\neg c$ mean that it is negative. Unfortunately, Mary has no prior knowledge about the a_i's or c. She has two sources of evidence—both equally and very reliable doctors who were involved in performing the test but who unfortunately completely disagree about the result of the test. Dr. Philippe says c is true; Dr. Pearl says $\neg c$ is true. Both of them would normally respect the other's opinion, but in this case they are aware of the disagreement, and each still strongly believes his own statement.

Take frame $\Omega = \Theta \times \{c, \neg c\}$, and for each i define b_i to be the proposition $a_i \Leftrightarrow c$ (i.e., b_i is true if and only if a_i and c have the same truth value), which corresponds to a subset of Ω.

Suppose that Mary is a Bayesian and that P is the probability function on Ω measuring her degrees of belief after receiving the two evidences. By symmetry $P(a_i) = P(a_j)$ for each $i, j = 1, \ldots, 1000$. Thus $P(a_1) = 0.001$.

It must be the case that $P(a_1 \mid b_1) = P(a_1 \mid \neg b_1)$, also by symmetry (this is so because given a_1 Dr. Philippe supports a_1 and Dr. Pearl supports $\neg a_1$, and given $\neg b_1$ Dr. Pearl supports a_1 and Dr. Philippe supports $\neg a_1$).The sandwich principle then implies that $P(a_1 \mid b_1) = P(a_1) = 0.001$.

This means that, after learning b_1, Mary is 99.9% certain that the patient does not have disease a_1 despite the fact that she was initially completely

ignorant about the truth of a_1, and of the two sources of evidence, one of them, Dr. Philippe, believes very strongly that the patient has a_1. Mary's degree of belief therefore seems completely unreasonable, as she is almost completely disregarding one of the two equally strong evidences; her certainty in $\neg a_1$ is quite unwarranted and could potentially be dangerous.

REMARKS (i) The same argument works also for other measures of belief. If P is a subadditive measure of belief obeying the sandwich principle, then we deduce that $P(a_1 \mid b_1) \leq 0.001$. This shows that the argument is against the sandwich principle rather than another argument against the Bayesian representation of ignorance (though they are connected; in Section 5.1 it is argued that the sandwich principle is incompatible with an adequate representation of ignorance).

(ii) The problem is symmetric with respect to the a_i's and also between b_1 and $\neg b_1$. Because of this it was assumed above that P should also have these symmetries. Though it would be extremely hard to justify a P without these symmetries, this assumption is not, in fact, essential. It must be case that for some i, $P(a_i) \leq 0.001$, and so by the sandwich principle $P(a_i \mid d) \leq 0.001$ for either $d = b$ or $d = \neg b$. The argument then proceeds as before.

(iii) It was not necessary to say how to represent the two evidences in a Bayesian model (if a larger frame were used, it must be a refinement of Ω, and the argument above still applies) or even to say exactly how strongly the doctors believed their statements—the argument above shows that, in this problem, the sandwich principle makes these issues irrelevant!

(iv) If the problem was simplified by not introducing c and the b_i's so that Dr. Philippe's evidence was just a_1 and Dr. Pearl's just $\neg a_1$, then Mary's posterior probability for a_1 would be sensible, about 0.5, since Dr. Philippe's evidence would be considered more surprising and thus get a higher likelihood ratio.

(v) It was not necessary to assume complete ignorance about the a_i's; it just makes the argument clearer. If Mary had some weak prior information, she should still not be 99.9% certain that Dr. Philippe's evidence is wrong.

(vi) A possible defense of Mary's degree of belief might be that Dr. Philippe on learning b_1 would radically change his mind about c. This would make sense if he originally had strong evidence implying that a_1 was false, but since he was initially ignorant about the truth of a_1, why should he change his mind much on learning that his statement implies a_1? After all, a_1 was entirely possible, and he was pretty certain that he observed c.

It is perhaps regrettable that the sandwich principle is not a principle of plausible reasoning. Suppose we use an adequate measure of belief for this problem, getting, say, $\text{Belf}(a_1 \mid b_1) = \text{Belf}(a_1 \mid \neg b_1) = 0.495$ (for example, if we used a Dempster–Shafer approach, representing the evidences of Drs. Philippe and Pearl by simple support functions with mass 0.98 assigned to c

and $\neg c$, respectively). By symmetry, $\text{Belf}(a_1 \mid b_i) = \text{Belf}(a_i \mid \neg b_i) = 0.495$ for all i.

Now suppose that, for each i, there is at Mary's disposal a test Test_i that determines if b_i is true or not. She can then choose (deliberately, randomly, or unconciously) any a_i to have highest belief after the next piece of evidence. By performing Test_i her belief in a_i will become 0.495, much higher than that for any other disease! This is not perhaps so strange—in life, our beliefs to some extent are determined by, and determine, the evidence we look for—but it does serve to remind one that there is a limit to how objective a degree of belief can be.

6. THE REPRESENTATION OF IF-THEN RULES

In Section 2.3 of his paper, Pearl criticizes the way a rule "**if** a **then** b **with certainty** α" is often represented in Dempster–Shafer theory, that is, by a simple support function with mass α attributed to the proposition $a \rightarrow b$. In this section it is shown how this representation of rules can be given a sound and natural interpretation. It is hard, then, to understand Pearl's claim (P.6) that Dempster–Shafer theory is not applicable for domains in which rules tolerate exceptions, including default reasoning; indeed, the limit of this Dempster–Shafer representation of rules is a special case of Reiter's default logic (Wilson [21]). Pearl suggests that rules should be interpreted as a conditional probability; there are, however, considerable problems with this interpretation, and, in any case, it does not seem credible that this is the only type of **if–then** rule.

6.1. An Interpretation of Dempster–Shafer Rules

In the introduction to his book on uncertainty, Pearl writes "uncertainty measures characterise invisible facts, i.e., exceptions not covered in the formulas" (Pearl [1, p. 2]). It then seems natural to represent an expert's rule **If** a **then** b: (α) by an underlying logical relationship $(n \wedge a) \rightarrow b$, where n is an unknown condition (or possibly just one that is hard to express) that the expert expects to be true α of the time.

As it is logically equivalent to $n \rightarrow (a \rightarrow b)$, this rule may also be interpreted as

> In a proportion α of worlds (or situations), we know that the material implication $a \rightarrow b$ is true.

But this conveys the same information as a simple support function with mass α attributed to $a \rightarrow b$, that is, the Dempster–Shafer representation of rules mentioned above. With a number of these rules, the combination using

Dempster's rule corresponds just to a lower probability given natural assumptions on the n's, the unknown conditions (Wilson [21]). For the purposes of this paper such a rule will be referred to as "the DS representation of rules," though other representations within Dempster–Shafer theory are possible.

6.2. Properties of DS Rules

Here the properties of DS rules that Pearl criticizes are explained: Contraposition one would normally expect, and chaining and reasoning by cases follow just from making the natural assumption that the unknown conditions of the two rules are independent, very similar to the assumption Pearl makes for noisy OR gates.

CONTRAPOSITION Because $n \wedge a \rightarrow c$ is logically equivalent to $n \wedge \neg c \rightarrow \neg a$, this type of rule allows contraposition.

I agree with Pearl that the contrapositive of a plausible rule is not necessarily plausible. For example, take the rule "Typically males don't have long beards." However, unless one is explicitly told otherwise, one would normally assume that a rule contraposes. To use Pearl's example (P.2.3.3), if all I am told is that "*If* a person is an orthodox Jew *then* that person refrains from eating pork, with certainty 0.999" and that Joe eats pork, then it seems natural to deduce that Joe is not an Orthodox Jew.[3]

CHAINING These rules also naturally chain; $n_1 \wedge a \rightarrow b$ and $n_2 \wedge b \rightarrow c$ with $\Pr(n_i) = \alpha_i$, $i = 1, 2$, leads to $n \wedge a \rightarrow c$, where $n \equiv (n_1 \wedge n_2)$, and $\Pr(n) = \alpha_1\alpha_2$ if we make the assumption that n_1 and n_2 are independent.

One would normally expect rules to chain. If one knew rules "**if** a **then** b" and "**if** b **then** c" and then learned a, one would usually deduce c. However, as Pearl points out (P.2.3.1) there are times when rules should not be allowed to chain; in particular, predictions should not trigger explanations, and so Dempster's rule should not be applied.

To say that the above pair of DS rules do not chain is just saying that n_1 and n_2 is an impossible event, or $\Pr(n_1 \wedge n_2) = 0$. Therefore, to calculate belief when there is a known procedure for determining if a pair of rules should not chain (or a list of all such pairs) requires only slight amendment to the usual simple and efficient Monte Carlo algorithm (Wilson [21, 22]): Suppose the rules **If** a_i **then** b_i: (α_i) for $i = 1, \ldots, m$ are known, and, for some proposition c, we want to find Bel(c). A large number of trials are performed.

[3] Indeed, logical rules not allowing contraposition have a rather odd property: If "**if** a **then** b" is such a rule and we learn $\neg b$, then we cannot deduce $\neg a$ despite the fact that if we learned a we would deduce a contradiction.

For each trial,

1. (a) For $i = 1, \ldots, m$
include i in σ with probability α_i;
(b) **If** (i) $\{a_i \rightarrow b_i: i \in \sigma\}$ is contradictory
or (ii) σ includes a pair of rules whose chaining should be suppressed
then Restart trial;

2. If $\{a_i \rightarrow b_i: i \in \sigma\}$ allows one to deduce c
then trial succeeds,
else trial fails.

The proportion of trials that succeed then converges to Bel(c).
The only difference with the usual algorithm is the addition of condition (ii).

REASONING BY CASES Given two rules **If** a **then** b: (α_1) and **If** $\neg a$ **then** b: (α_2), which we will interpret as uncertain material implications $n_1 \wedge a \rightarrow b$ and $n_2 \wedge \neg a \rightarrow b$ with $\Pr(n_i) = \alpha_i$, $i = 1, 2$. Again assuming independence of the n_i's we get $\Pr(b) \geq \Pr(n_1 \wedge n_2) = \alpha_1\alpha_2$, so $\text{Bell}(b) = \alpha_1\alpha_2$.

The reason the lower probability Bel(b) is as low as this is that in worlds where $n_1 \wedge \neg n_2$ is true, b may be always false if a is always false; in the event $\neg n_1 \wedge n_2$, b may be always false if a is always true; and in the event $\neg n_1 \wedge \neg n_2$ there is no constraint on b so b may again always be false—in this case $\Pr(b)$ actually equals $\alpha_1\alpha_2$.

A Bayesian interpretation of **if–then** rules may well be in terms of likelihood ratios (see, e.g., Pearl [23]), but any such representation is also supposedly "in clear violation of common sense" (P.2.3.3) since given rules **If** a **then** b: (0.99) and **If** $\neg a$ **then** b: (0.99) it could be the case that $\Pr(b)$ is small if the prior probability of b was very small.

Those familiar with Pearl's noisy OR gates may have noticed the close connection; exception independence, where "Each exception to normal behaviour acts as an independent variable" (Pearl [1, p. 185]) is essentially the assumption used above to deduce the chaining and reasoning by cases behavior that Pearl criticizes.[4]

There will be many cases where the use of the simple independence assumptions between the n_i's is not appropriate. Arbitrary probability functions over the n_i's could instead be used, representing correlations between rules, a simple case of which is suppressing chaining as described above. The Monte Carlo algorithm can again be used, with modification of step 1, to calculate the values of Bel (which will still be a belief function) and will be efficient if the representation of the probability function is efficient. For example, if the probability function (i.e., the correlation information about the

[4] Pearl himself is aware of the connection with noisy OR gates: see Pearl [1, p. 446]). Exception independence is one of the two assumptions of noisy OR gates, where there are two causal conditions of a variable.

rules) is represented by a Bayesian network (Pearl [1]), then a single trial of a Monte Carlo algorithm on this network could be used for step 1, to generate σ.

Another example of where Dempster's rule should not be used is in Pearl's reasoning by cases examples (P.2.3.3). These would be better represented by the use of a constant belief function. In the first, the set of three rules would be represented by a belief function with $\text{Bel}[(M \to O) \wedge (J \to O) \wedge (C \to O)] = 0.001$, $\text{Bel}[(M \to O) \wedge (J \to O)] = 0.7$, and $\text{Bel}(M \to O) = 0.9$, M standing for Muslim, etc.

6.3. Conditional Probability Interpretations of If-Then Rules

Pearl suggests a conditional probability interpretation of rules: **If** a **then** b: (α) gets interpreted as a conditional probability $\Pr(b \mid a) = \alpha$; indeed, in Pearl [1, p. 450], he appears to go further when talking about default rules (i.e., logical representations of uncertain rules) when he says that the "proper interpretation" of rules is by conditional probability statements with probabilities close to 1. This extreme position seems rather implausible when one considers the many different types of rules that philosophers and logicians have invented.

There are a number of problems with conditional probability interpretation of rules:

1. *Rule does not adequately represent support.* If an expert tells us a rule, **If** a **then** b: (0.3), intending to give weak support to the rule, she would be horrified to then discover that it had been interpreted as $\Pr(\neg b \mid a) = 0.7$, equivalent to the rule **If** a **then** $\neg b$: (0.7).
2. *Consistency.* If we learn two rules **If** a **then** b: (0.6) and **If** a **then** b: (0.62), then the knowledge base is inconsistent, and similarly if we learn from one expert that **If** a **then** b: (0.7) and from another that **If** a **then** $\neg b$: (0.4), then again inconsistency results. This is a problem caused by the constraints being hard, as opposed to the soft constraints of the DS representation (see Section 6.4).
3. *Contraposition.* One would normally expect a rule to contrapose, as pointed out in Section 6.2.1, but this type of rule does not naturally do so.
4. *Chaining.* Take two rules

If my alarm clock wakes me up **then** I'll catch my train (0.99), and
If I catch my train **then** I won't be late for work (0.99).

Given that my alarm clock does wake me up, it seems natural to deduce that I probably won't be late for work. However, using a conditional probability interpretation of these rules, one cannot deduce anything about the probability that I won't be late for work.

5. *Irrelevant information nullifies rules.* Given the rule **If** Orthodox-

Jew(x) **then** Observe(x): (0.999), on learning new facts Orthodox-Jew(joe) and Has-Sister(joe), we cannot deduce anything about whether joe observes or not.

The fourth and fifth problems, and perhaps the third, mean that some sort of "independence" or "irrelevance" assumptions must be made (then maybe use a lower probability approach, given these assumptions and the constraints generated by the rules) to make such an interpretation of rules practical. For the limiting case, where the certainties of the rules tend to 1 (Adams's logic of conditionals), Geffner and Pearl have made progress, defining an "irrelevance" predicate (Geffner [24], Pearl [1, p. 493]), but it is not easy to see how such an approach could be generalized for the usual, noninfinitesimal, case.

An alternative idea, also explored by Pearl [1, p. 491] for the limiting case, is to use maximum entropy to pick a single probability function given the constraints. As indicated in Section 3.1, it is not clear to me that this approach can be guaranteed to give sensible results.

6.4. Different Types of Rules

Other representations of rules have been suggested, representing the rule **If** a **then** b: (α) by $\Pr(b \mid a) \geq \alpha$ and by $\Pr(a \rightarrow b) \geq \alpha$. These also suffer from problems mentioned above, though they are both (especially the former) also plausible representations of certain types of rules.

An important difference between these rules and DS rules is that the latter treat a rule as a "soft constraint," so that if a rule is contradicted to some degree by other rules then its reliability is reduced. For example, fact a and DS rule **If** a **then** b: (0.4) lead to $\Pr(b) \geq 0.4$, and so $\mathrm{Bel}(b) = 0.4$, but adding an extra rule **If** a **then** $\neg b$: (0.5) would reduce the reliability of the first rule, thus reducing $\mathrm{Bel}(b)$.

The other representations mentioned treat rules as hard constraints—the reliability of a rule is not changed by learning other rules that contradict it. This condition is a convenient but strong condition, it treats rules as absolute rather than as objects that should be conditioned on other information received.

There are clearly many representations of rules. For example, (i) representations that use hard rather than soft constraints and (ii) representations based upon making the material implication uncertain, which allow contraposition, and those based upon conditional probability, which do not always allow contraposition. There are also (iii) representations in which rules act as supports, so that given fact a and two such rules of the form **If** a **then** b: (0.3), the belief assigned to b will be higher than 0.3. No single representation can be appropriate for all situations; instead, frameworks should be developed that can represent many different types. The DS rule, although representing a natural type of rule computationally efficiently, is, like the conditional probability representation favored by Pearl, necessarily limited in its representational power.

7. DEMPSTER-SHAFER THEORY AS RANDOMIZED LOGIC

Pearl's logic-based view of Dempster–Shafer theory is enlightening (P.1.4). Artificial intelligence has recently seen an explosion in the development and use of nonstandard logics (see, e.g., Smets et al.[25], Ginsberg [26]). The additivity of conventional probability makes it very hard to sensibly define a probability function over logics, but uncertainty can be added to just about any logic in a natural and very simple way using a Dempster–Shafer approach (Wilson [21]).

Imagine a system accumulating items of knowledge a_i, $i = 1, \ldots, m$, that are propositions in some logic, a_i coming from source S_i. In a straight logical approach the source is assumed to be completely reliable, and so a knowledge base $K = \{a_1, \ldots, a_m\}$ is built up. It is often desirable to be able to add degrees of certainty to the facts and rules in a knowledge base, so instead of assuming that each source S_i is completely reliable, assume that each source has probability α_i of being reliable; thus if E_i is the event that it is reliable, then $E_i \Rightarrow a_i$, and our prior probability of E_i is α_i.

In the straight logical case every proposition c was either deducible from the knowledge base K or not. Now, since the sources have only a probability α_i of being reliable, each a_i is known only with probability α_i, and so c is deducible with only a certain probability. Bel(c) is defined to be the probability that c is deducible from the knowledge base; Bel is, as Pearl eloquently puts it, the probability of provability.

One could think of the knowledge base no longer consisting of a fixed set of propositions, but of random propositions, each a_i being present only α_i of the time, so that other propositions will be deducible only a fraction of the time, this fraction being their belief. This view leads to a Monte Carlo algorithm for calculating belief.

8. CONCLUSIONS

No single measure can capture all one's intuitions about belief. Although in many ways an attractive measure, Bayesian belief (as well as other measures that obey the sandwich principle) can exhibit counterintuitive behavior in situations of partial ignorance. Alternative models of belief, such as Dempster–Shafer belief and lower probability, which cope better with such situations, need further exploration. In particular, more examples of their practical application are desirable.

I argue in my reply to Glenn Shafer's paper (elsewhere in this issue) that not all belief function representations of evidence are appropriate for combination by Dempster's rule. For example, Bayesian belief functions in Dempster–Shafer theory cannot represent a Bayesian prior adequately, and this

is the cause of the counterintuitive results in examples such as the three prisoners puzzle. The use of Dempster–Shafer belief functions to update a Bayesian prior (P.6) seems to me not to have been adequately justified. I believe that Pearl is right in suggesting that, although it represents certain types of evidence very nicely, the representational power of Dempster–Shafer theory is limited, and ways of mixing Dempster–Shafer representations with other representations of evidence should be looked into.

Just as there are many plausible models of belief, none of which corresponds to all one's intuitions, there are many types of if–then rules, including the DS representation and conditional probability, no single representation being always completely adequate. Also when using the theory in a rule-based or logic-based system, Dempster's rule should not be blindly applied—care must be taken to represent any dependencies.

ACKNOWLEDGMENTS

Many colleagues and friends have taken the time to discuss ideas expressed in this paper with me. I am especially indebted to Mike Clarke, David Heckerman, Ita O'Keeffe, Judea Pearl, Philippe Smets, David Spiegelhalter, and Bill Triggs.

References

1. Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, Los Altos, Calif., 1988.
2. Lauritzen, S. L., and Spiegelhalter, D. J., Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *J. Roy. Stat. Soc., Ser. B* **50**(2), 157–224, 1988.
3. Bonissone, P., and Henrion, M., Eds., *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, MIT, Cambridge, Mass., 1990.
4. Clark, D. A., Barton, G. J., and Rawlings, C. J., A knowledge-based architecture for protein sequence analysis and structure prediction, *J. Mol. Graphics*, **8** (June), 49–107, 1990.
5. Gordon, J., and Shortliffe, E. H., A method of managing evidential reasoning in a hierarchical hypothesis space, *AI* **26**, 323–357, 1985.
6. Pearl, J., On evidential reasoning in a hierarchy of hypotheses, *AI* **28**, 9–15, 1986.
7. Shafer, G., Probability judgment in artificial intelligence and expert systems (with discussion), *Stat. Sci.* **2**(1), 3–44, 1987.

8. Jaynes, E. T., Where do we stand on maximum entropy? in *The Maximum Entropy Formalism* (R. D. Levine and M. Tribus, Eds.); MIT Press, Cambridge, Mass., 1979.

9. Paris, J. B., and Vencovská, A., On the applicability of maximum entropy to inexact reasoning, *Int. J. Approx. Reasoning* **3**(1), 1–33, 1989.

10. Paris, J. B., and Vencovská, A., A note on the inevitability of maximum entropy, *Int. J. Approx. Reasoning* 1987.

11. Smith, C. A. B., Consistency in statistical inference and decision, *J. Roy. Stat. Soc., Ser. B* **23**: 218–258, 1961.

12. Nilsson, N. J., Probabilistic logic, *AI* **28**(1), 71–87, 1986.

13. van der Gaag, L., Computing probability intervals under independency constraints, in [3], 491–497.

14. Kyburg, H. E., Jr., Bayesian and non-Bayesian evidential updating *AI* **31**, 271–293, 1987.

15. Fagin, R., and Halpern, J. Y., A new approach to updating beliefs, in [3], pp. 317–325.

16. Dempster, A. P., Upper and lower probabilities induced by a multivalued mapping, *Ann. Math. Stat.* **38**: 325–339, 1967.

17. DeCampos, L. M., Lamata, M. T., and Moral, S., The concept of conditional fuzzy measure, *Int. J. Intell. Syst.* 5, 237–246, 1990.

18. Jaffray, J.-Y., Bayesian updating and belief functions, Res. Rep. Laboratoire d'Informatique de la Décision Université de Paris VI, July 1990.

19. Shafer, G., and Tversky, A., Languages and designs for probability, *Cogn. Sci.* **9**, 309–339, 1985.

20. Simpson, E. H., The interpretation of interaction in contingency tables, *J. Roy. Stat. Soc., Ser. B* **13**: 238–241, 1951.

21. Wilson, N., Rules, belief functions and default logic, in [3], pp. 443–449.

22. Wilson, N., Justification, computational efficiency and generalisation of the Dempster–Shafer theory, Res. Rep. 15, June 1989, Dept. of Computing and Mathematical Sciences, Oxford Polytechnic.; also to appear in *AI*.

23. Pearl, J., How to do with probabilities what people say you can't, *Second Conf. on AI Applications: The Engineering of Knowledge-Based Systems*, Miami Beach, Fla., Dec. 11–13, IEEE Computer Soc. Press, Washington, D.C., 6–12, 1985.

24. Geffner, H., Default reasoning: causal and conditional theories, Ph.D. Thesis, Computer Science Dept. UCLA, Los Angeles, Calif., November 1989.

25. Smets, P., Mamdami, E., Dubois, D., and Prade, H., *Non-standard Logics for Automated Reasoning*, Academic, London, 1988.

26. Ginsberg, M. L., Ed., *Readings in Non-Monotonic Reasoning*, Morgan Kaufmann, Los Altos, Calif., 1987.